ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

A Survey on Data Stream and Its Various Techniques

Hiral Desai, Dharmik Vasiyani, Jay Gandhi

PG Student, Dept. of Computer Engineering, B.H.Gardi College Of Engineering & Technology, Rajkot, India

PG Student, Dept. of Computer Engineering, B.H.Gardi College Of Engineering & Technology, Rajkot, India

Assistant Prof., Dept. of IT, B.H.Gardi College Of Engineering & Technology, Rajkot, India

ABSTRACT:Data StreamMining is become new emerging topic for research in knowledge discovery. In this continuous changing nature of data creates problem in mining the knowledge from it and its difficult to store. There are some techniques and algorithms which are using for mining in the data stream like classification, clustering and frequent patterns. Here gives an overview of all these techniques with their merits and demerits.There are mainly four challenges termed as Concept drift, Infinite length, Concept evaluation and Limited labeled data. Here also gives description of all these challenges.With reference to all these problem here gives discussion of future work.

KEYWORDS: data stream, concept drift, concept evolution, infinite length, limited labeled data.

I. Introduction

For the last decades, algorithms of data mining have been proposed for the knowledge discovery from the huge amount of data. But these types of algorithms we can apply only on static datasets. Those are not giving the exact output in the case of dynamic datasets and rapidly incrementing data.

Continuous flow of data is known as data stream. It likes as river continuously flow in and out. It is difficult to store and mining. Computer network traffic, web searches, web click ATM transactions, Phone calls conversations are all examples of stream data.

We summarize the characteristics of data streams as follows:

- Enormous volumes of continuous data.
- Rapidly changing and real time and requires quick response.
- Because of limited storage only the summary of the data can be stored .
- The data arriving is multidimensional and of low level so the techniques to mine such data needs to be very sophisticated.

Data stream mining is a process of extracting valuable knowledge structure from unremitting and immediate data records. There are mainly two types: Online Stream mining and Offline Stream Mining. Network traffic, credit card detection and many other real time applications are examples of online stream mining. And generating report based on web log streams is example of offline streams[1].

Data streams have intrinsic properties which make it troublesome for the customary data mining systems to order stream data. Some of the most challengingproperties of data streams include but not limited to infinite length, conceptdrift, concept evolution, limited labeled data. Classification and Clustering are two most important techniques to deal with data stream. . All these three techniques applies various algorithm to improve the accuracy of stream data mining and decrease the space complexity.

Infinite length

Data streams are thought to be infinite in length. Along these lines, unreasonable to store it in main memory and utilize all the recorded data for training[1][2]. In view of that customary multi-pass learning algorithms are not straightforwardly material to data streams. One of the Proposed arrangement is to gap stream into equal sized chunk.

DOI: 10.15680/IJIRCCE.2015. 0309090

Concept Drift

Data streams perceive concept-drift, which occurs when the essential concept of the data changes over time. In order to address concept-drift, a classification model must continuously update itself to the most recent concept.In other words we can say, Concept drift means that the statistical properties of the target variables which the model is trying to predict, change over time in unforseen ways, this causes problems because the predictions become less accurate as time passes.Examples of real life concept drifts include monitoring systems, financial fraud detection spam categorization weather predictions and customer preferences[5][8][9][10].

Deviations of object concepts are categorized into sudden, incremental, gradual, recurring, blip or noise drifts. Figure 1.Shows the six basic types of drifts. The principal plot (Sudden) demonstrates unexpected changes that right away and irreversibly change the information cases of individual class. The following two plots (Incremental and Gradual) represent changes that happen gradually after some time. Incremental drift occurs when data example slowly change their values over time, and gradual drift occurs when the change in data example includes the class distribution of various data[2].

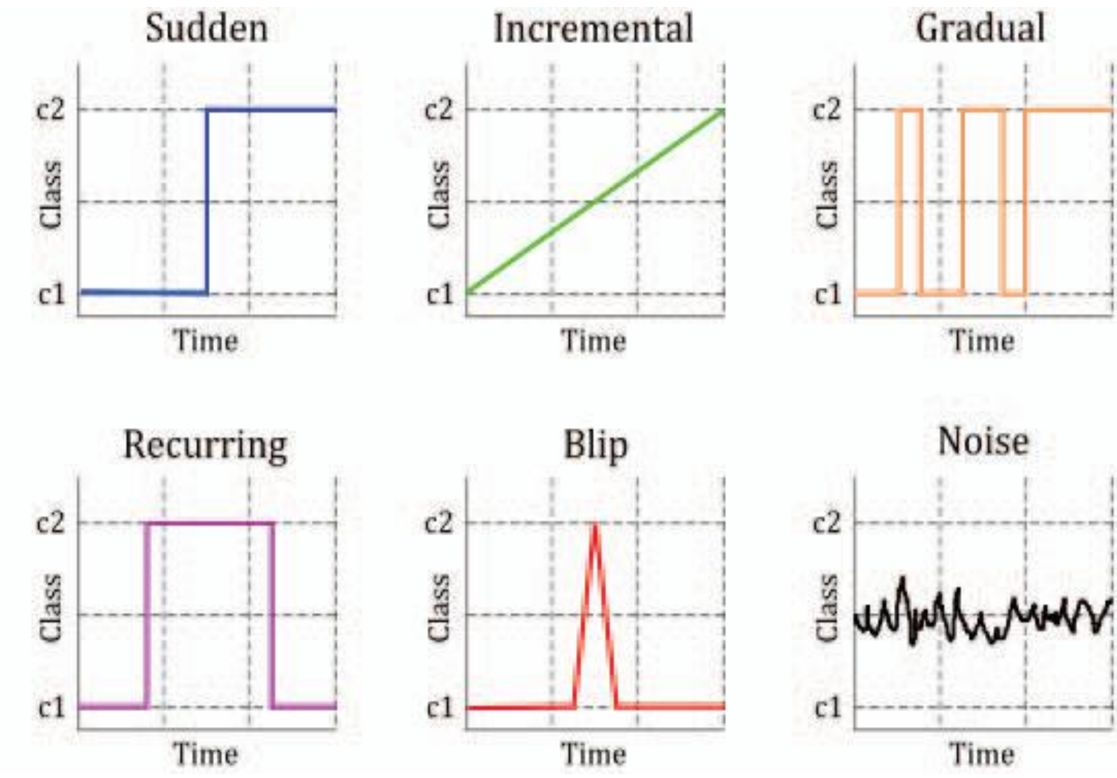

Figure. 1 [2]

Based on the number of classifier we can divide the classifier in two categories. 1) Single Classifier Approach 2) Ensemble Approach. Single classifier approach handle concept drift using one classifier like Decision Tree, Naïve Bayesian, Hoeffding Tree, Adaptive Hoeffding tree, ADWIN, DDM, Flora, CVFDT etc. Whereas Ensemble Approach uses two or more classifier to handle concept drift like Streaming Ensemble Algorithm, Accuracy Weighted Ensembles, Hoeffding Option Trees etc[3].

Limited labelled data

In data streams, data arrives very quickly. The speed at which data points are labeled tags far behind the speed at which at data points arrive in the stream[8][9][12].The annotations of labeled samples are frequently time consuming and sometimes impossible to acquire in many real-world problems. So, a good classifier in streaming environment should be able to defer the traininguntil true labels become available yet continuing labeling newly arrived instancesusing the current classifier[2]. Moreover, it should be able to use partially labeledtraining dataActive learning and semi supervised learning have been proposed as an alternative approach to solve limited labeled data which jointly exploit labeled and unlabeled samples for training classifiers to expanding classification accuracy.Classification and clustering based on active learning methods such as Support Vector method, Bayesian rule and neural network. Some Semi supervised learning models are self-training, mixer models, graph based methods, co-training, multiview learning for limited labeled data[10].

Concept Evolution

Information streams additionally recognize concept-evolution, which happens when a novel class shows up in the stream. So as to handle with concept-evolution, a classification model must have the capacity to mechanically recognize novel classes when they show up, before being trained with the marked instances of the novel class[3][11].

The remainder of this paper is organized as follows. In Section 2 discussed the related work of data stream techniques. In Section 3 discussed comparative analysis of various techniques. Finally, in Section 4discussed the most interesting conclusions.

II. RELATED WORK

In this paper [3], propose a various methods which deals with the concept drift. FLORA framework, Meta learning method, CD algorithms, fuzzy information network method, decision tree methods, ensemble classifying methods including expert ensemble classifier method, ensemble classifier method upcoming the basic classifier set and ensemble classifier method with multi algorithm. Treatment of the concept drift , noise and unlabeled data stream, luation standards of classifier's performance, selection of characteristic attributes and adaptive adjustment of training window's size all are the major challenging research are in classification of data stream.

Dayrelis Mena and et.al. introduced a new techniquefor achieve good performance in novel class detection named as a Similarity based data stream Classifier (SimC)[4]. It also used for remove useless instance vale that not add to classification process. Author compared SimC with other well known methods. Based on accuracy, SimC algorithm is very competitive. This algorithm uses the advantages of the instance-based learning techniques. It divides in mainly three procedures: Build classifier, update the classifier, classify the new instance.

Dariusz Brzezinski and Jerzy Stefanowski proposedanother incremental ensemble classifier named as an Online Accuracy Updated Ensemble which gave best normal classification accuracy in view of three general procedures, for example, Online Component assessment, presentation of an incremental learner and the utilization of a float identifier for changing a piece ensemble into an incremental learner[5]. OAUE calculation keeps up a weighted arrangement of segment classifier and predicts the class of aggregating so as to approach illustrations the expectations of segments utilizing a weighted voting tenet. It utilizes the hoeffding trees as part classifiers.

StARMiner Tree (ST) propose an incremental decision tree algorithm which is based on very fast decision tree (VFDT) technique to mining medical data and also propose a decision tree model constructed from numerical data using statistics to decide when to perform the division of tree node. ST is based on StARMiner (Statistical Association Rule Miner) algorithm which is used to mine association rulesover continuous feature values[6].

This paper [7], present two partitioning clustering calculation in particular CLARANS and ECLARANS (enhanced CLARANS) are utilized for clustering and identifying the exceptions in data streams . Taking into account two execution variables, for example, clustering exactness and exception location precision. ECLARANS calculation perform all the more precisely. Clustering is one of the unsupervised methodology in data stream.

Xindong Wuand et.al. presented a Semi-supervised classification calculation for data streams with concept drifts and UNlabeled data (SUN)[8]. In this calculation, Authors fabricate a decision tree incrementally and produce concept clusters at leaves in a grouping calculation created from k-Modes. With the system of base up hunt and the deviation of classification utilizing concept clusters, they recognize concept drifts from clamor. Exploratory studies uncover the productivity and viability of SUN even in the cases with a vast volume of unlabeled data. SUN is productive and compelling. It could track concept drifts well when there is a huge volume of unlabeled data.

A REDLLA algorithm which is a Semi-supervised classification algorithm for data streams. REDLLA remains for Recurring concept Drifts and Limited LAbeled data in which a choice tree is embraced as the classification model. At the point when growing a tree, a clustering algorithm in light of k-Means is introduced to deliver concept clusters and mark unlabeled data at takes off. In perspective of deviations between history concept clusters and new ones, potential concept drifts are recognized and repeating concepts are kept up. Broad studies on both engineered and genuine data affirm the benefits of our REDLLA algorithm more than two best in class online classification algorithms and a few known online semi-supervised algorithms, even for the situation with more than 90% unlabeled data[9].

A new semi-supervised ensemble learning (SSEL) algorithm proposed in [10] for the classification of streaming data. The approach could be categorized as modified self-training but without the conventional problems of self-training. In self-training methods, if the learners are weak and predict the label of unlabeled instances incorrectly, using the unlabeled data will degrade the performance of the learner. But in SSEL showed that if the number of instances in each window is enough, then the algorithm is PAC learnable and noise will not degrade the performance of the learner. This paper used the incremental decision tree as the base classifier of ensemble learners.Decision trees are appropriate classifiers for online learning as they are fast and accurate learners.

In this paper [11], creators proposed structure named as an ActMiner, which stands for Active Classifier for Data Streams with novel class Miner, performs classification and novel class location in data streams while obliging little measure of labeled data for preparing. ActMiner coordinates the answers for four noteworthy data stream classification issues: infinite length, concept-drift, concept evolution, and limited labeled data. ActMiner reduces data requiring so as to mark time and expense just a couple chose cases to be labeled. Indeed, even with this limited measure of labeled data, it outperforms cutting edge data stream classification strategies that utilization ten times or more labeled data. Algorithms and a few known online semi-supervised algorithms, even for the situation with more than 90% unlabeled data[11].

Authors of [12] proposed an algorithm for online data stream classification and learning with limited labels using selective self-training semi supervised classification.To incrementally learn from both labeled data and unlabeled data and selection of the data to be trained can be done as soon as the classification process is complete by Selective self-training method. Method divides into three parts: off line pre-training, online classification and cluster reduction.The capacity of the proposed strategy to gain from limited labels is demonstrated by using so as to accomplish 95% normal precision just 1% labeled information.

Dino Iencoand et.al. proposed another dynamic learning methodology for developing information streams in light of pre-clustering stride, for selecting the most informative occurrences for marking. Proposed calculation ACLStream (Active Clustering Learning for Data Streams)[13]. This methodology abuses a clustering based partitioning of the information space to center examining on the possibly most valuable cases to mark. Bunches are positioned by homogeneity of their anticipated class circulations. Occurrences in every group are positioned by variables: most extreme a posteriori classification likelihood, and geometrical position inside the bunch. Sureness is characterized to be the result of these two elements. Examples with low conviction inside a given group are favored, as they speak to main issues over which the classifier is more questionable.

Author propose an ensemble classification method based on a decision- feedback[14]. The fundamental rule of the methodology is making full utilization of both labeled and unlabeled occurrences in the data stream to enhance the performance.At first, data streams are divided intoproper-sized data chunks and single classification model istrained from each chunk. An ensemble model E consists ofthese models, the instances in an unlabeled chunk areroughly predicted firstly by E. Unsupervised models are prepared from these unlabeled instances to give useful requirements. With the limitations as input data, the instances labels are ordered all the more accurately by fulfilling the agreement maximization of E and these unsupervised models.

III. COMPARATIVE ANALYSIS

Method	Approach	Demerits	Merits
Graph based method	Semi supervised	Clear numerical system. Performance is solid if the graph happens to fit the undertaking . The (pseudo) opposite of the Laplacian can be seen as a portion network . Can be stretched out to directed graphs.	Performance is evil if the graph is wicked. Thoughtful to graph construction and edge weights.
Co-training	Semi supervised	Humblebindingtechnique. Applies to every single existing classifies.	✓Natural feature splits may not exist. Models using both features should do better.
S3VM	Semi supervised	Pertinent wherever SVMs are appropriate. Clear mathematical System.	Optimization challenging. Can be trapped in poor nearby optima. More humble supposition than generative model or graph-based methods, possibly lesser addition.
Hoeffding option tree	Ensemble	Memory used with tree expansion. Number of candidate attributes. Could spend a lot time with ties.	Can deal with extremely large datasets. Each example to be read at most once in a small constant time. Makes it possible to mine online data sources. Build very complex trees with acceptable computational cost.
Self-Training	Semi supervised	Promptinaccuracies could emphasize themselves. But, there are exceptional situations when self-training is proportional to the Expectation-Maximization (EM) calculation.	The humblest semi-supervised learning method. A covering method, applies to existing (complex) classifiers. Frequently used in actual tasks like natural language processing.

IV. CONCLUSION

Data stream is continuous flow of data and rapidly increments. Due to thisrapid increment in data, storage of stream data is become difficult task.And also mining some pattern and discovering knowledge from it is problematic task. In this paper , presented some challenging issues in data stream mining and given an overview of stream data mining techniques. Infinite length, concept drift, novel class detection and limited labeled data are main challenges in data stream mining . To solve concept drift and limited labeled data some techniques are discussed here. As future work, the plan to carry out active learning semi supervised approach to solve the problem of limited labeled data in data stream for reducing the size of storage and also reducing the time and cost of assigning the label to data stream.Active learning approach used for solving the problem of limited labeled data and ensemble learning approach suitable for concept drift.

REFERENCES

1. Darshana Parikh1, Priyanka Tirkha 'In Data Streams Using Classification And Clustering Different Techniques To Find Novel Class', International Journal Of Research In Engineering And Technology (IJRET) Volume: 02 Issue: 08, 2013.
2. Ms. Priyanka B.Dongre, Dr. Latesh G. Malik,'A Review On Real Time Data Stream Classification And Adapting To Various Concept Drift Scenarios', International Advance Computing Conference (IACC) ,IEEE,2014.
3. Ouyang Zhenzheng , Gao Yuhai et.al. 'Study On The Classification Of Data Streams With Concept Drift', Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), IEEE, 2011.
4. Dayrelis Mena-Torres, Jesús S. Aguilar-Ruiz, 'A Similarity-Based Approach For Data Stream Classification' Elsevier-Expert Systems With Applications 41, pp. 4224–4234, 2014.
5. Dariusz Brzezinski, Jerzy Stefanowski, 'Combining Block-Based And Online Methods In Learning Ensembles From Concept Drifting Data Streams', Elsevier-Information Sciences 265,pp. 50–67, 2014.
6. Mirela Teixeira Cazzolato And Marcela Xavier Ribeiro 'A Statistical Decision Tree Algorithm For Medical Data Stream Mining', CBMS,IEEE,2013.
7. Dr. S. Vijayarani, Ms.P.Jothi 'Partitioning Clustering Algorithms For Data Stream Outlier Detection', International Journal Of Innovative Research In Computer And Communication Engineering (IJRCCE) Vol. 2, Issue 4, 2014.
8. Xindong Wu, Peipei Li , Xuegang Hu, 'Learning from concept drifting data streams with unlabeled data', Elsevier- Neurocomputing 92, pp. 145–155, 2012.
9. Xindong Wu, Peipei Li , Xuegang Hu , 'Mining Recurring Concept Drifts with Limited Labeled Streaming Data' JMLR: Workshop and Conference Proceedings 13: pp. 241-252 , 2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, pp. 8-10, 2010.
10. Zahra Ahmadi and Hamid Beigy, 'Semi-supervised Ensemble Learning of Data Streams in the Presence of Concept Drift' Springer-Verlag Berlin Heidelberg , pp. 526–537, 2012.
11. Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham , 'Classification and Novel Class Detection in Data Streams with Active Mining', Advance in knowledge discovery and data mining, volume. 6119, pp. 311-324, 2010.
12. Loo Hui Ru, Trias Andromeda, M. N. Marsono, 'Online Data Stream Learning And Classification With Limited Labels', Proceeding Of International Conference On Electrical Engineering, Computer Science And Informatics (EECSI) Yogyakarta, Indonesia , 2014.
13. Dino Ienco, Albert Bifet, Indr˙e ˇZliobait˙e, and Bernhard Pfahringer, 'Clustering Based Active Learning For Evolving Data Streams', Springer-verlag Berlin Heidelberg, pp. 79-93, 2013.
14. LIU Jing, XU Guo-sheng, ZHENG Shi-hui, XIAO Da, GU Li-ze 'Data streams classification with ensemble model based on decision-feedback' Elsevier 21, pp. 79-85,.

BIOGRAPHY

Hiral Desai is a ME student in the Computer Engineering Department in B.H. Gardi college of engineering & Technology, Rajkot, Gujarat, India. She received Bachelor of Information Technology degree in 2014 from L.E. College, Morbi, Gujarat, India. Her research interests are Data Mining, Big data and Artificial intelligence etc.

Dharmik Vasiyani is a ME student in the Computer Engineering Department in B.H. Gardi college of engineering & Technology, Rajkot, Gujarat, India. He received Bachelor of Information Technology degree in 2014 from L.E. College, Morbi, Gujarat, India. Her research interests are Data Mining and Security etc.

Jay Gandhi is an Assistant Professor of Information Technology Department in B.H. Gardi college of engineering & Technology, Rajkot, Gujarat, India. He received Master of Technology in information Technology from Charusat University, Changa.